LAPPEENRANNAN-LAHDEN TEKNILLINEN YLIOPISTO LUT
School of Engineering Science
Laskennallisen tekniikan koulutusohjelma
Kandidaatintyö

Annika Vieraankivi

Markkinointidatan tutkiminen muuttujanvalintamenetelmää ja luokittelualgoritmia käyttäen

Ohjaaja: Pasi Luukka

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto

School of Engineering Science

Laskennallisen tekniikan koulutusohjelma

Annika Vieraankivi

Markkinointidatan tutkiminen muuttujanvalintamenetelmää ja luokittelualgoritmia käyttäen

Kandidaatintyö

2019

27 sivua, 5 kuvaa, 10 taulukkoa

Ohjaaja: Pasi Luukka

Avainsanat: muuttujanvalinta; luokittelu; epämääräisyys; samankaltaisuus;

Tämän kandidaatintyön tavoitteena oli tehdä markkinointidatasta ennustemalli, jolla voidaan ennustaa, kannattaako tietylle kohderyhmälle lähteä markkinoimaan. Oleellista oli myös löytää aineistosta ongelman kannalta merkittävät muuttujat. Muuttujanvalinnan avulla voitiin yksinkertaistaa mallia ja nopeuttaa laskentaa.

Työssä esiteltiin muuttujanvalinnan ja luokittelun teoriat. Muuttujanvalinnassa keskityttiin epämääräisyysarvoon perustuvaan menetelmään ja hyödynnettiin sitä datan esikäsittelyssä. Luokittelualgoritmina käytettiin similaarisuuteen pohjautuvaa menetelmää. Tuloksia arvioitiin luokittelun keskitarkkuuden, -spesifisyyden ja -herkkyyden kautta.

Löydettiin ennustemalli, joka suoriutui hyvin kymmentä muuttujaa käyttäen. Luokittelutuloksia onnistuttiin parantamaan muuttujanvalinnan avulla. Luokittelun keskitarkkuus oli 83,97%, keskispesifisyys 86,83% ja keskiherkkyys 69,69%. Markkinointikampanjaa kuvaavat muuttujat olivat tulosten perusteella ongelman kannalta tärkeämpiä kuin asiakkaaseen liittyvät muuttujat.

Sisältö

1 JOHDANTO

Kandidaatintyön tavoitteena on pyrkiä ennustamaan, kannattaako tietylle kohdeyleisölle lähteä markkinoimaan. Tämä tehdään datan perusteella, mistä pyritään tekemään ennustemalli, joka pystyisi ennustamaan tuon mahdollisimman hyvin. Ennustemallin luomiseen käytetään muuttujanvalintamenetelmää sekä luokittelualgoritmia.

1.1 Tausta

Suurien datamäärien kasvaessa datan monet ulottuvuudet ovat nousseet tärkeäksi tutkimuskohteeksi. Datamäärien kasvaessa luokittelutulokset voivat parantua huomattavasti, mutta suuremmaksi ongelmaksi on noussut tärkeiden muuttujien tunnistaminen ja luokittelun kannalta merkityksettömien tai tarpeettomien muuttujien poistaminen. Ylimääräiset muuttujat häiritsevät datan opettamista ja kasvattavat algoritmin suoritusaikaa [4]. [13]

Moro, Laureano ja Cortez [15] tutkivat kyseistä dataa CRoss-Industry Standard Process for Data Mining eli CRISP-DM-menetelmällä. Tarkoituksena oli löytää malli, joka selittää markkinoinnnin onnistumisen. Työssä tutkittiin False Positive Raten ja True Positive Raten suhdetta. Heidän tulosten perusteella vektoritukikoneella (engl. vector support machine) luotu ennustemalli suoriutui parhaiten vertailluista malleista.

Tan, Sim ja Yeoh [17] tutkivat samankaltaista dataa liittäen korrelaatiota hyödyntävän muuttujanvalintamenetelmän (CFS) ja osajoukkojen yhtenäisyyteen (SC) liittyvän muuttujanvalintamenetelmän ennustemalliin. Ennustemalleja verrattiin muihin samankaltaisiin malleihin. Heidän tulokset osoittivat, että datan esikäsittely paransi mallin suoriutumista.

1.2 Ennustemallin luominen

Datan tutkiminen suoritetaan esikäsittelyn ja luokittelun kautta. Esikäsittelyn avulla voidaan tutkia eri muuttujien merkityksellisyyttä ja parantaa luokittimen suoriutumista. Luokittelun tarkoituksena on rakentaa mahdollisimman hyvä ennustemalli kohderyhmälle markkinoimisesta.

Tuloksien avulla voidaan tutkia luokittelun suoriutumista ja merkitseviä muuttujia. Luokittelutarkkuuden perusteella nähdään, kuinka hyvin luokitin onnistuu luokittelemaan tapauk-

set. Luokittelutarkkuuden muuttumisen perusteella muuttujanvalinnan jälkeen voidaan tehdä päätelmiä siitä, kuinka tärkeitä tietyt muuttujat ovat. Virheluokkien perusteella voidaan arvioida virheen vakavuutta, sillä eri virheluokkaan kuuluvat näytteet ovat ongelman kannalta vakavampia. Ongelman kannalta vakavin virheluokka on false negative, eli sellaisten asiakkaiden, jotka olisivat tilanneet talletuksen, luokittelu negatiiviseen luokkaan. Virhe on vakavin siksi, että mallin mukaan tällaisille asiakkaille ei kannata lähteä markkinoimaan. On haitallisempaa menettää asiakkaita, jotka olisivat tilanneet talletuksen kuin markkinoida turhaan asiakkaille, jotka eivät tilaa talletusta.

Ongelmaa rajattiin valitsemalla luokittelualgoritmiksi similaarisuusarvoon perustuva menetelmä. Menetelmä valittiin sillä perusteella, että haluttiin tutustua tiettyyn luokittelualgoritmiin ja sen matemaattiseen taustaan tarkemmin.

Työn tarkoituksena oli myös tutustua epämääräisyysarvoon perustuvaan muuttujanvalintamenetelmään. Toinen yleisesti käytössä oleva aihepiiri on piirteiden poiminta (engl. feature extraction), jonka tarkoituksena on rakentaa alkuperäisestä joukosta uusi muuttujien osajoukko [1]. Erona muuttujanvalintaan, piirteiden poiminnassa luodaan uusia muuttujia. Piirteiden poiminta voidaan toteuttaa esimerkiksi pääkomponenttianalyysillä [**dunteman1989principal**].

1.3 Työn rakenne

Tämä kandidaatintyön raportti sisältää kirjallisuuskatsauksen erilaisia muuttujanvalintamenetelmiä ja luokittelualgoritmeja käsittelevistä töistä sekä markkinointidataa tutkivista tutkimuksista. Sen lisäksi käsitellään tutkimusmetologian osalta muuttujanvalintamenetelmiä sekä luokittelualgoritmeja ja niiden osalta olennaista matemaattista taustaa.

Raportissa esitellään työssä käytettävät laskentatavat sekä laskennassa käytetty materiaali. Laskennan tulokset ja niiden analysointi esitetään omassa kappaleessaan. Lopuksi pohditaan tuloksia ja aiheen mahdollisia ongelmakohtia.

2 ONGELMAN KUVAUS

Työn tavoitteena on tehdä ennustemalli markkinointidatan pohjalta. Datassa on 16 eri muuttujaa, jotka kuvastavat asiakkaan tietoja. Tiedot liittyvät sekä asiakkaan henkilökohtaisiin tietoihin, että markkinointikampanjaan ja aikaisempiin kampanjoihin liittyviin seikkoihin. Lisäksi datassa on tieto siitä, onko asiakas hyväksynyt tarjouksen, eli onko markkinointi onnistunut.

2.1 Työn tavoitteet

Työssä pyritään ennustamaan, onko tietylle kohdeyleisölle markkinoiminen kannattavaa. Oleellista on myös selvittää, mitkä muuttujat ovat merkittäviä ongelman kannalta.

2.2 Työn toteutus

Ennustemallin tekemisessä käytetään epämääräisyysarvoon perustuvaa muuttujanvalintamenetelmää. Algoritmin avulla poistetaan yksi kerrallaan vähiten merkitseviä muuttujia datasta ja tarkastellaan sen vaikutuksia luokittelutarkkuuteen, -herkkyyteen ja -spesifisyyteen. Luokittelualgoritmina käytetään similaarisuusmittaan perustuvaa menetelmää ja tuloksia tutkitaan eri simialarisuusparametreilla.

Laskenta suoritetaan tekemällä luokittelu ennen muuttujien poistoa, sekä jokaisen muuttujan poistamisen jälkeen. Luokittelutuloksia analysoidaan ja tutkitaan myös eri muuttujien merkitystä mallissa. Tutkitaan löydetyn ennustemallin tuloksia tarkemmin ja vertaillaan sitä alkutilanteeseen.

2.3 Aineisto ja ohjelmistot

Ennustemalli tehdään csv-muotoisen aineiston perusteella [15]. Aineisto liittyy pankkien suoramarkkinointiin Portugalilaisissa pankeissa. Markkinointikampanja perustui puhelinsoittoihin. Aineisto sisältää 16 muuttujan verran tietoa pankin asiakkaista ja 17. muuttuja kertoo, onko asiakas tilannut määräaikaistalletuksen eli onko markkinointi onnistunut. Taulukossa 1 on esitetty aineiston muoto. Osa muuttujista on numeerisia ja osa kategorisia. Koko datassa on 45211 tapausta, ja datasta on tehty pienempi versio, jossa on satunnaisesti valittu

10% koko aineiston tapauksista. Pienempi versio datasta on tarkoitettu suurempaa laskentaa vaativien testien suorittamiseen. Aluksi datasta poistetaan puuttuvat arvot ja muutetaan data numeeriseksi. Toinen vaihtoehto olisi ollut korvata puuttuvat arvot. Laskenta suoritetaan Matlab-ohjelmistolla.

Asiakkaasen liittyvät muuttujat ovat ikä, työn tyyppi (ylläpito, ei tiedossa, työtön, hallinnollinen, kodinhoitaja, yrittäjä, opiskelija, haalarityöntekijä, itsensä työllistävä, eläköitynyt, tekniikka, palvelut) siviilisääty (naimisissa, eronnut, naimaton), koulutus (ensimmäinen, toinen tai kolmas aste), onko luottoa, keskimääräinen vuotuinen saldo, onko asuntoa, onko lainaa, miten asiakasta on lähestytty (ei tiedossa, puhelimitse, matkapuhelimitse), viime kontaktin päivä, viime kontaktin kuukausi, viime kontaktin kesto, kampanjan aikaisten yhteydenottojen määrä, viime kampanjasta kuluneiden päivien määrä (-1, jos asiakasta ei ole ennen lähestytty), kontaktien määrä ennen kampanjaa ja viime kampanjan lopputulos (ei tiedossa, muu, epäonnistuminen, onnistuminen).

Taulukko 1. Esimerkki datasta

age	30	33	35	30	59
job	unemployed	services	management	management	blue-collar
marital	married	married	single	married	married
education	primary	secondary	tertiary	tertiary	secondary
default	no	no	no	no	no
balance	1787	4789	1350	1476	0
housing	no	yes	yes	yes	yes
loan	no	yes	no	yes	no
contact	cellular	cellular	cellular	unknown	unknown
day	19	11	16	3	5
month	oct	may	apr	jun	may
duration	79	220	185	199	226
campaign	1	1	1	4	1
pdays	-1	339	330	-1	-1
previous	0	4	1	0	0
poutcome	unknown	failure	failure	unknown	failure
y	no	no	no	no	no

3 TUTKIMUSMETOLOGIA

Ratkaisussa hyödynnetään muuttujanvalintamenetelmää sekä luokittelualgoritmia. Tässä kappaleessa tutustutaan tarkemmin aihealueisiin ja käytettyihin menetelmiin.

3.1 Muuttujanvalinta

Muuttujanvalinta (engl. feature selection) on menetelmä, jolla poistetaan epäolennaisia muuttujia, jotka rasittavat tehtävän suorittamista. Muuttujanvalinnan tarkoituksena on auttaa ymmärtämään dataa, vähentää laskennan määrää, vähentää ulottuvuuksien määrää ja parantaa ennusteen suoriutumista [7]. Tarkoituksena on siis löytää kyseessä olevan ongelman kontekstissa olennainen muuttujien osajoukko [4]. Poistamalla tarpeettomat muuttujat informaatio voidaan saavuttaa vähemmällä määrällä muuttujia, jotka sisältävät maksimiaalisen tiedon luokista. Merkityksettömät muuttujat eivät liity ratkaisun kontekstiin, ja tarpeettomat muuttujat eivät tuo ratkaisuun mitään uutta [8]. Muuttujanvalintamenetelmät voidaan jakaa kolmeen pääluokkaan, sulautettuihin menetelmiin (engl. embedded) ja suodatin- (engl. filter) sekä kääremenetelmiin (engl. wrapper) [2]. Suodatinmenetelmät käsittelevät muuttujia ennen luokittelua, kun taas sulautetut menetelmät ja kääremenetelmät käsittelevät muuttujanvalintaa osana luokittelua [7]. Tässä työssä käytetään suodatinmenetelmää.

Kääremenetelmien pääidea on se, että ennustetta tarkastellaan ”mustana laatikkona”, ja ennusteen suoriutumisen perusteella valitaan muuttujien joukko. Täten ne ovat erityisen yleiskäyttöisiä ja yksinkertaisia [7]. Muuttujien osajoukot etsitään peräkkäisillä valinta-algoritmeilla tai heuristisilla hakualgoritmeilla. Molemmat tavat johtavat sisäkkäisiin muuttujien osajoukkoihin [7]. Peräkkäisten valinta-algoritmien suorittaminen aloitetaan tyhjällä joukolla, johon lisätään muuttujia yksitellen, kunnes funktion suurin hyöty on saavutettu. Tätä kutsutaan peräkkäishauksi (engl. sequential forward selection) [7]. Vaihtoehtoisesti voidaan aloittaa käyttäen kaikkia muuttujia, josta niitä poistetaan yksi kerrallaan, eli peräkkäinen muuttujien vähentäminen (engl. sequential backward selection). Heuristiset hakualgoritmit arvioivat erilaisia muuttujista koostuvia osajoukkoja. Erilaisia hakumenetelmiä ovat esimerkiksi best-first -haku [11], branch-and-bound -hakualgoritmi [16] ja geneettiset algoritmit [6]. Kääremenetelmien suurin ongelma on niiden laskennallinen vaativuus. Jokaiselle osajoukolle on luotava oma ennustemalli, ja suurien data-aineistojen tapauksessa ennusteen opettaminen voi viedä suurimman osan algoritmin suoritukseen kuluvasta ajasta.

Suodatinmenetelmät järjestelevät muuttujat tietyn kriteerin perusteella. Niiden pääidea on

suodattaa tehtävän kannalta turhat muuttujat ennen luokittelua. Muuttujat järjestellään ja kynnysarvon alapuolelle jäävät muuttujat poistetaan. Yleisiä kriteerejä, joilla muuttujat voidaan järjestellä on esimerkiksi Pearsonin korrelaatiokriteeri [7], muuttujan kyky erotella näytteet [9], jokaisen muuttujan ja kohteen välisen informaation tarkastelu [18] ja Fisherin kriteeri [5]. Tässä työssä muuttujat järjestellään niiden epämääräisyysarvon (engl. nonspecificity) ja similaarisuusarvon (engl. similarity) perusteella. Suodatinmenetelmien etuja ovat niiden yksinkertaisuus ja nopeus johtuen siitä, että menetelmä vaatii vain kriteerin mukaisen arvon laskemisen jokaiselle muuttujalle ja niiden järjestelyn. [2] [7]

Sulautetuissa menetelmissä muuttujanvalinta suoritetaan datan opettamisprosessin aikana [7]. Sulautetutuilla menetelmillä yritetään tarjota ratkaisua suodatinmenetelmien ja kääremenetelmien ongelmille. Niiden tarkoituksena on vähentää eri osajoukoilla luokitteluun kuluvaa aikaa. Ratkaisu löydetään nopeammin, kun ennustetta ei opeteta uudelleen jokaisella osajoukolla. Dataa käytetään tehokkaammin, sillä opetusdataa ei tarvitse jakaa sekä opetus- että validointikäyttöön. Päätöspuut [3] [12] on yleisesti käytetty sulautettu muuttujanvalintamenetelmä. [7]

Digitalisaation myötä data-aineistot kasvavat ja merkitsevien muuttujien määrän epävarmuus samalla kasvaa. Tämän vuoksi suurien data-aineistojen tapauksissa on otettava huomioon laskentaan kuluva aika. Kääremenetelmät ovat usein tarkempia, mutta laskennallisesti huomattavasti raskaampia, kun taas suodatinmenetelmät ovat nopeampia, minkä vuoksi ne soveltuvat paremmin suurien aineistojen käsittelyyn.

3.1.1 Matemaattinen tausta

Epämääräisyys (engl. nonspecificity) liittyy vaihtoehtoisten osajoukkojen kardinaliteetteihin, eli alkioiden lukumäärään. Se on määritelty joukko-opissa seuraavasti [10]:

$$U(A) = c \cdot \log_b |A|$$

jossa $|A|$ kuvaa joukon A kardinaliteetteja ja b ja c ovat positiivisia vakioita. Esimerkiksi, jos $b = 2$ ja $c = 1$, epävarmuus (engl. uncertainty) lasketaan bitteinä ja saadaan

$$U(A) = \log_2 |A|$$

Funktiota U kutsutaan Hartleyn funktioksi. Kun Hartleyn funktio liitetään joukon X osajoukkoihin, niin $U : \mathscr{P}(X) - \{\emptyset\} \to \mathbb{R}^+$ välillä $0 \leq U(A) \leq \log_2 |X|$.

Hartleyn funktion antama epävarmuus riippuu joukosta A. Määrän väheneminen voidaan laskea käyttäen $I(A,B)$. Tämä on yhtä suuri kuin vähentynyt epävarmuus (engl. reduced

uncertainty), joka saadaan erotuksesta $U(A) - U(B)$. Tämä voidaan kirjoittaa myös muotoon

$$I(A,B) = \log_2 \frac{|A|}{|B|}$$

Jos esimerkiksi määrätään $|B| = 1$, niin $I(A,B) = \log_2 |A| = U(A)$. $U(A)$ on siis joukon A yhden alkion luonnehdintaan tarvittavan informaation määrä.

Similaarisuusarvon avulla voidaan vertailla datanäytteitä. [19] mukaan similaarisuus määritellään seuraavasti:

$$s(x,y) = 1 - |x - y|$$

jossa $x, y \in [0,1]$. Similaarisuus voidaan myös esittää [14] mukaan:

$$s(x,y) = (1 - |x^p - y^p|)^{1/p}$$

jossa $p \in \mathbb{R}^+$ voidaan käyttää similaarisuuden vahvuuden säätämisessä.

3.1.2 Epämääräisyysarvoon perustuva muuttujanvalinta-algoritmi

Seuraavaksi esitellään Luukan ja Lohrmannin [13] esittelemä epämääräisyysarvoon perustuva muuttujanvalinta-algoritmi. N merkitsee luokkien määrää, n näytteiden määrää, n_i näytteiden määrää luokassa i ja D muuttujien määrää.

1. Normalisoidaan data välille [0,1]

$$X^D \to [0,1]^D$$

2. Lasketaan keskiarvovektorit v_i testinäytteille

$$v_{i,d} = (\frac{1}{n_i} \sum_{x \in x_i} X_d), \quad d = 1,\ldots,D, \quad i = 1,\ldots,N$$

jossa x_d merkitsee muuttujaa d ja keskiarvo lasketaan näytteille, jotka kuuluvat luokkaan i.

3. Lasketaan similaarisuusarvot $S(x_{j,d}, v_{i,d})$, jossa $x_j \in X_i$, näytevektoreiden x_j ja keskiarvovektoreiden v_i välillä. Muuttujalle d samankaltaisuus saadaan:

$$S(x_{j,d}, v_{i,d}) = (1 - |(x_{j,d}^P - (v_{i,d})^P|)^{\frac{1}{P}}$$

$$d = 1,\ldots,D, \quad i = 1,\ldots,N, \quad j = 1,\ldots,n$$

4. Lasketaan epämääräisyysarvo jokaiselle muuttujalle d luokassa i hyödyntäen kohdassa 3 laskettuja similaarisuusarvoja.

$$u_{1,i,d} = \ln(\sum_{x_j \in x_i} S(x_{j,d}, v_{i,d}))$$

5. Lasketaan similaarisuusarvot $S(x_{j,d}, v_{i,d})$, jossa $x_j \in X_i$ kaikkien keskiarvovektoreiden v_i ja näytteiden x_j välillä. Muuttujalle d samankaltaisuus saadaan:

$$S(x_{j,d}, v_{i,d}) = (1 - |(x_{j,d}^P - (v_{i,d})^P|)^{\frac{1}{P}}$$

6. Lasketaan epämääräisyysarvo jokaiselle muuttujalle d luokassa i hyödyntäen kohdassa 5 laskettuja similaarisuusarvoja.

$$u_{2,i,d} = \ln(\sum_{x_j \in x} S(x_{j,d}, v_{i,d}))$$

7. Lasketaan vähentynyt epävarmuus jokaiselle muuttujalle

$$T_d = \sum_{i=1}^{N} (u_{1,i,d}/n_i - u_{2,i,d}/n)$$

8. Poistetaan datasta muuttujat, joilla on kynnysarvoa suurempi epävarmuuden muutos

Seuraavaksi esitellään yksinkertainen esimerkki algoritmin käytöstä. Taulukossa 2 on mallidata, jossa on viisi vektorinäytettä, joilla jokaisella on 3 muuttujaa (f_1, f_2, f_3). Luokan tunniste C on neljännessä sarakkeessa.

Taulukko 2. Yksinkertainen esimerkkidata

f_1	f_2	f_3	C
1.6	0.4	2.5	1
2.0	0.2	4	1
1.1	1	0.8	1
0.4	3	2.2	2
0.5	2.8	4.3	2

1. Normalisoidaan data välille [0,1]. $X^D \rightarrow [0,1]^D$. Tämä on esitetty taulukossa 3.

Taulukko 3. Normalisoitu data

f_1	f_2	f_3	C
0.8333	0.1875	0.6471	1
1	0.1250	0.9412	1
0.6250	0.3750	0.3137	1
0.3333	1	0.5882	2
0.3750	0.9375	1	2

2. Lasketaan keskiarvovektorit v_i testinäytteille

$$v_1 = [0.8194, 0.2292, 0.6340]$$

$$v_2 = [0.3542, 0.9688, 0.7941]$$

3. Lasketaan similaarisuusarvot $S(x_{j,d}, v_{i,d})$, jossa $x_j \in X_i$, näytevektoreiden x_j ja keskiarvovektoreiden v_i välillä. Nämä on esitetty taulukossa 4.

Taulukko 4. Keskiarvovektoreiden ja luokkien väliset similaarisuudet

Similaarisuus	f_1	f_2	f_3
$S(x_{1,d}, v_{1,d})$	0.9861	0.9583	0.9869
$S(x_{2,d}, v_{1,d})$	0.8194	0.8958	0.6928
$S(x_{3,d}, v_{1,d})$	0.8056	0.8542	0.6797
$S(x_{4,d}, v_{2,d})$	0.9792	0.9688	0.7941
$S(x_{5,d}, v_{2,d})$	0.9792	0.9688	0.7941

4. Lasketaan epämääräisyysarvo jokaiselle muuttujalle d luokassa i hyödyntäen kohdassa 3 laskettuja similaarisuusarvoja. Nämä ovat taulukossa 5.

Taulukko 5. Epämääräisyysarvot u_1

Epämääräisyys	f_1	f_2	f_3
$u_{(1,1,d)}$	0.9598	0.9963	0.8584
$u_{(1,2,d)}$	0.6721	0.6614	0.4626

Jossa esimerkiksi muuttujalle f_1 epämääräisyys on $u_{1,1,1} = ln(0.9861 + 0.8194 + 0.8056) = 0.9598$.

5. Lasketaan similaarisuusarvot $S(x_{j,d}, v_{i,d})$, jossa $x_j \in X_i$ kaikkien keskiarvovektoreiden v_i ja näytteiden x_j välillä. Nämä on esitetty taulukossa 6.

Taulukko 6. Keskiarvovektoreiden ja näytteiden väliset similaarisuudet

Similaarisuus	f_1	f_2	f_3
$S(x_{1,d}, v_{1,d})$	0.9861	0.9583	0.9869
$S(x_{2,d}, v_{1,d})$	0.8194	0.8958	0.6928
$S(x_{3,d}, v_{1,d})$	0.8056	0.8542	0.6797
$S(x_{4,d}, v_{1,d})$	0.5139	0.2292	0.9542
$S(x_{5,d}, v_{1,d})$	0.5556	0.2917	0.6340
$S(x_{1,d}, v_{2,d})$	0.5208	0.2188	0.8529
$S(x_{2,d}, v_{2,d})$	0.3542	0.1563	0.8529
$S(x_{3,d}, v_{2,d})$	0.7292	0.4063	0.5196
$S(x_{4,d}, v_{2,d})$	0.9792	0.9688	0.7941
$S(x_{5,d}, v_{2,d})$	0.9792	0.9688	0.7941

6. Lasketaan epämääräisyysarvo jokaiselle muuttujalle d luokassa i hyödyntäen kohdassa 5 laskettuja similaarisuusarvoja. Nämä on esitetty taulukossa 7.

Taulukko 7. Epämääräisyysarvot u_2

Epämääräisyys	f_1	f_2	f_3
$u_{(2,1,d)}$	1.3031	1.1722	1.3731
$u_{(2,2,d)}$	1.2705	1.0002	1.3386

7. Lasketaan vähentynyt epävarmuus jokaiselle muuttujalle

$$T = [0.1413, 0.2283, -0.0249]$$

Tässä esimerkiksi $T_1 = 0.9598/3 - 1.3031/5 + 0.6721/2 - 1.2705/5 = 0.1413$

8. Poistetaan datasta muuttujat, joilla on kynnysarvoa suurempi epävarmuuden muutos. Poistetaan siis muuttuja f_3.

3.2 Luokittelu

Luokittelun tarkoituksena on luokitella näytteet luokkiin mahdollisimman tarkasti. Muuttuja-avaruus jaetaan alueisiin, ja jokaiselle luokalle valitaan yksi alue. [13]

Tässä työssä käytetään similaarisuusarvoon perustuvaa luokittelumenetelmää [14]. Menetelmä perustuu näytteiden ja luokkien similaarisuusasteiden vertaamiseen. Datamatriisi jaetaan N määrään luokkia ja jokaiselle luokalle pyritään etsimään ideaalivektori (engl. ideal vector). Ideaalivektoreiden on tarkoitus kuvastaa luokkaa mahdollisimman hyvin. Luokan i ideaalivektori merkitään $v_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,D})$. Ideaalivektorien määrittely:

$$v_{i,d} = (\frac{1}{n_i} \sum_{x \in x_i} {X_d}^m)^{\frac{1}{m}}, \quad \forall d = 1, \ldots, D, \quad i = 1, \ldots, N$$

jossa n_i kuvaa näytteiden määrää luokassa i. Vektoreiden $x_j \in X$ luokkien määrittämiseksi niitä verrataan kaikkien luokkien ideaalivektoreihin. Näytteen x ja luokan ideaalivektorin v_i välinen similaarisuus:

$$S\langle x_j, v_i \rangle = (\frac{1}{D} \sum_{d=1}^{D} (1 - |{x_{j,d}}^p - {v_{i,d}}^p|)^m)^{\frac{1}{m}}$$

Näyte $x_j \in X$ kuvaa luokkaa, jolla on suurin similaarisuusarvo, esimerkiksi:

$$Luokka(\mathbf{x_j}) = \mathbf{arg} \max_{i=1,\ldots,N} \mathbf{S}\langle \mathbf{x_j}, \mathbf{v_i} \rangle$$

Luokittelun suorituskykyä voidaan mitata erilaisilla mittareilla. Näistä yleisimmin käytetty on tarkkuus (engl. accuracy), joka määritellään [1] mukaan:

$$Tarkkuus = \frac{oikein\ luokitellut\ tapaukset}{kaikki\ tapaukset}$$

Luokittelutulokset voidaan jakaa neljään luokkaan, true positive (TP), false positive (FP), true negative (TN) sekä false negative (FN) [1]. Tässä työssä true positive -luokkaan kuuluvat asiakkaat, joille markkinointi on onnistunut ja tapaus on luokiteltu oikein. True negative -luokkassa ovat asiakkaat, jotka eivät ole tilanneet markkinoitua talletusta, ja ovat luokiteltu oikein. True negative -luokassa ovat asiakkaat, jotka ovat luokiteltu siten, että markkinointi ei toiminut, ja se on todellisuudessa oikein. False negative -ryhmään kuuluvat asiakkaat, jotka on luokiteltu siten, että markkinointi ei toiminut, vaikka todellisuudessa se toimi. Näiden luokkien perusteella voidaan laskea erilaisia mittareita siitä, kuinka hyvin luokittelu on onnistunut.

Luokittelun herkkyys (engl. sensitivity) kuvaa sitä, kuinka hyvin luokitin tunnistaa tapaukset, jotka ovat positiivisia. [1] mukaan herkkyys määritellään seuraavasti:

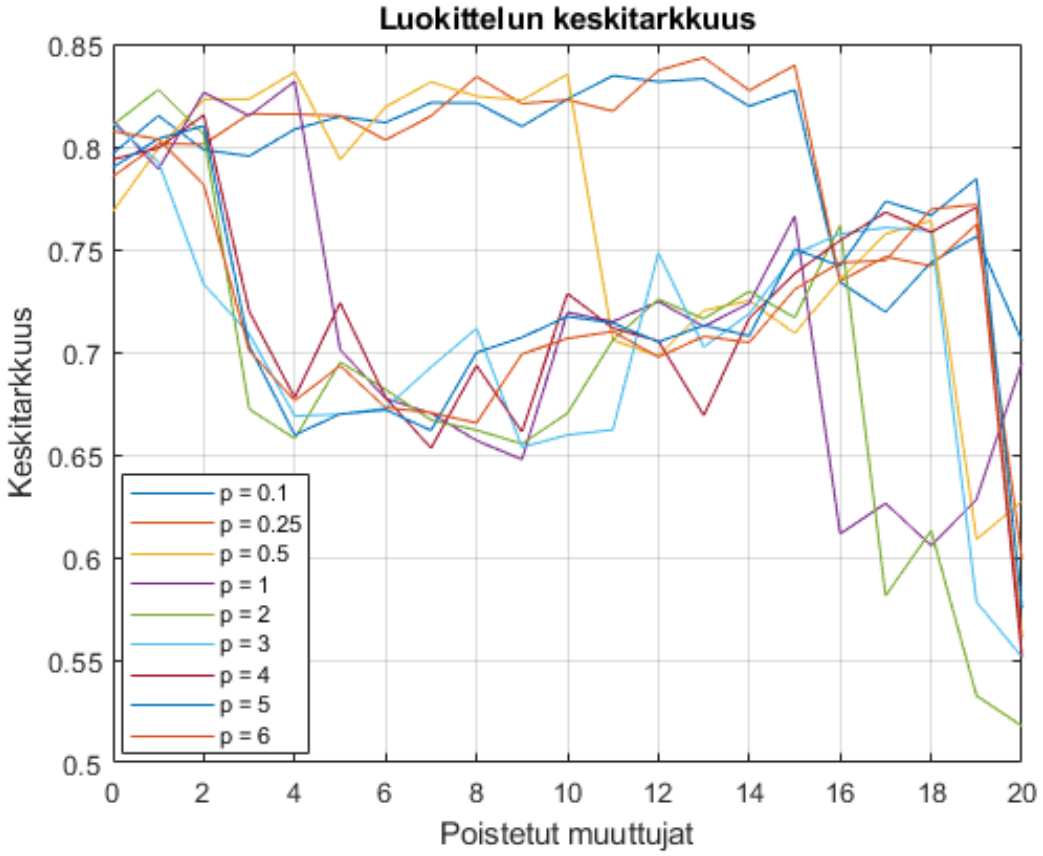
$$Herkkyys = \frac{TP}{TP+FN}$$

Luokittelun spesifisyys (engl. specificity) liittyy luokittimen kykyyn tunnistaa tapaukset, jotka ovat negatiivisia. [1] mukaan spesifisyys määritellään:

$$Spesifisyys = \frac{TN}{TN+FP}$$

4 TULOKSET

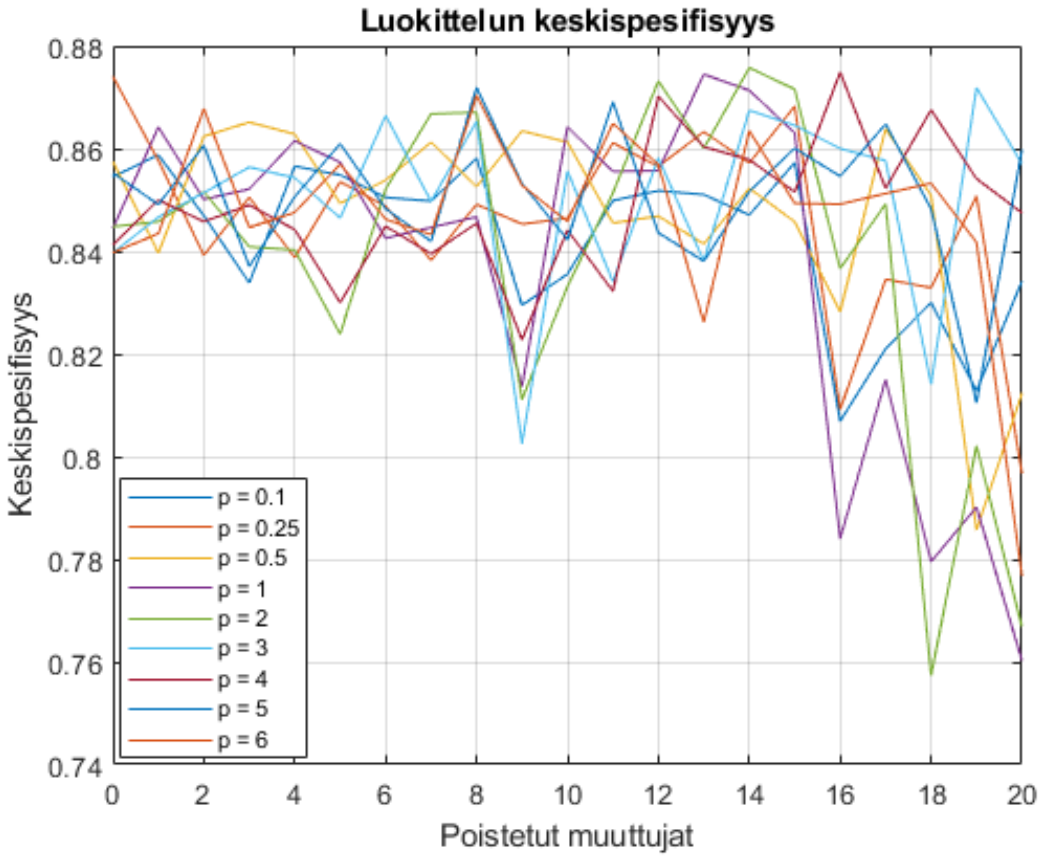
Laskenta toteutettiin tekemällä ensiksi luokittelu kaikkia muuttujia käyttäen. Data jaettiin 70% opetukseen ja 30% testaukseen. Opetusdatasta 50% käytettiin opetukseen ja 50% validointiin. Data jaettiin 30 kertaa sattumanvaraisesti testaus- ja opetusdataksi. Yksi kerrallaan poistettiin muuttujanvalinnalla vähiten merkittävä muuttuja, jonka jälkeen luokittelu tehtiin uudelleen. Joka luokittelun keskitarkkuudesta, -spesifisyydestä ja -herkkyydestä tehtiin kuvaajat, jossa luokittelun suoriutumista kuvaava mitta on pystyakselilla ja poistettujen muuttujien määrä vaaka-akselilla.



Kuva 1. Luokittelun keskitarkkuus muuttujia poistettaessa eri similaarisuusparametreilla

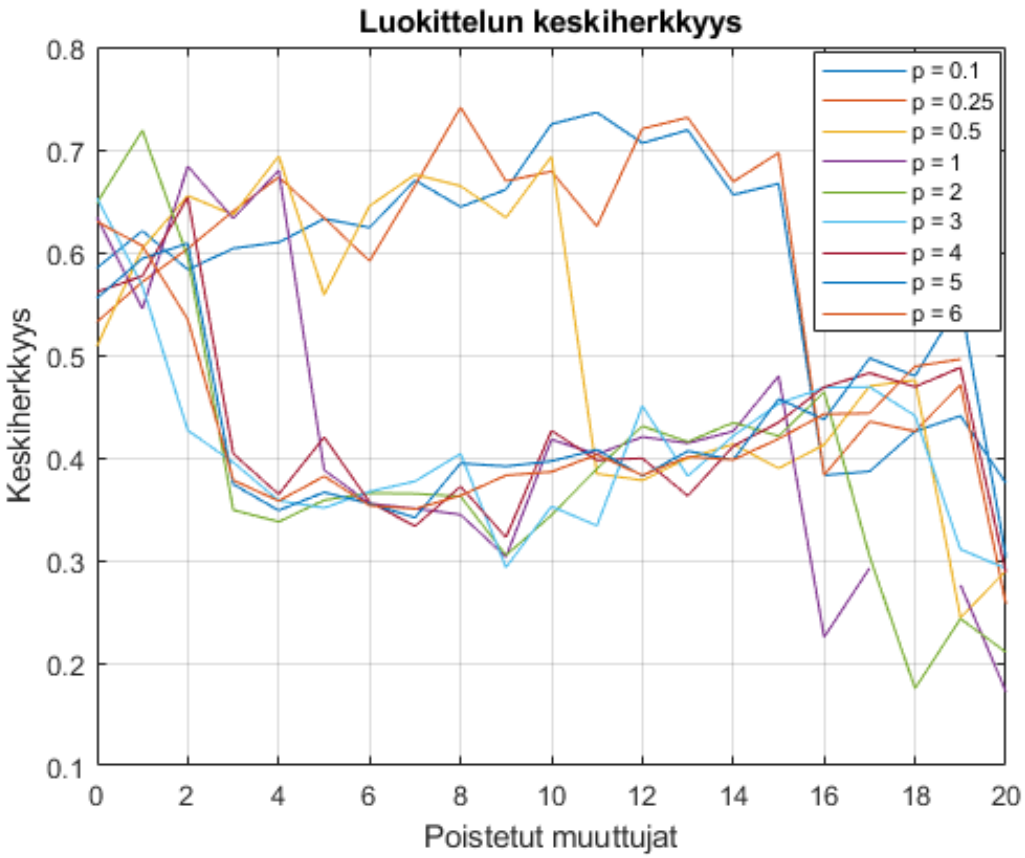
Kuvassa 1 on esitetty luokittelun keskitarkkuudet eri similaarisuusparametreilla. Kuvaajasta voidaan nähdä, että similarisuusparametrin arvoilla p=0,1 ja p=0,25 luokittelutulos paranee 15:nnen muuttujan poistamiseen saakka. Ennustemallia voidaan siis yksinkertaistaa ja nopeuttaa poistamalla muuttujat, jotka eivät tuo lisäarvoa ennusteelle. Korkein saatu tarkkuus on 84,34%, joka saavutetaan 13:nnen muuttujan poistamisen jälkeen arvolla p=0,25. Parhaana tuloksena voidaan pitää p=0,25 arvolla saatua 83,97% 15:n muuttujan poistami-

sen jälkeen, sillä arvot eivät eroa tilastollisesti merkitsevästi toisistaan 1% riskitasolla. Similaarisuusparametreilla 2-6 keskitarkkuuden arvot ovat hyvin samankaltaisia, erona kuvaajan loppuvaiheessa tapahtuvan romahduksen paikka.



Kuva 2. Luokittelun keskispesifisyys muuttujia poistettaessa eri similaarisuusparametreilla

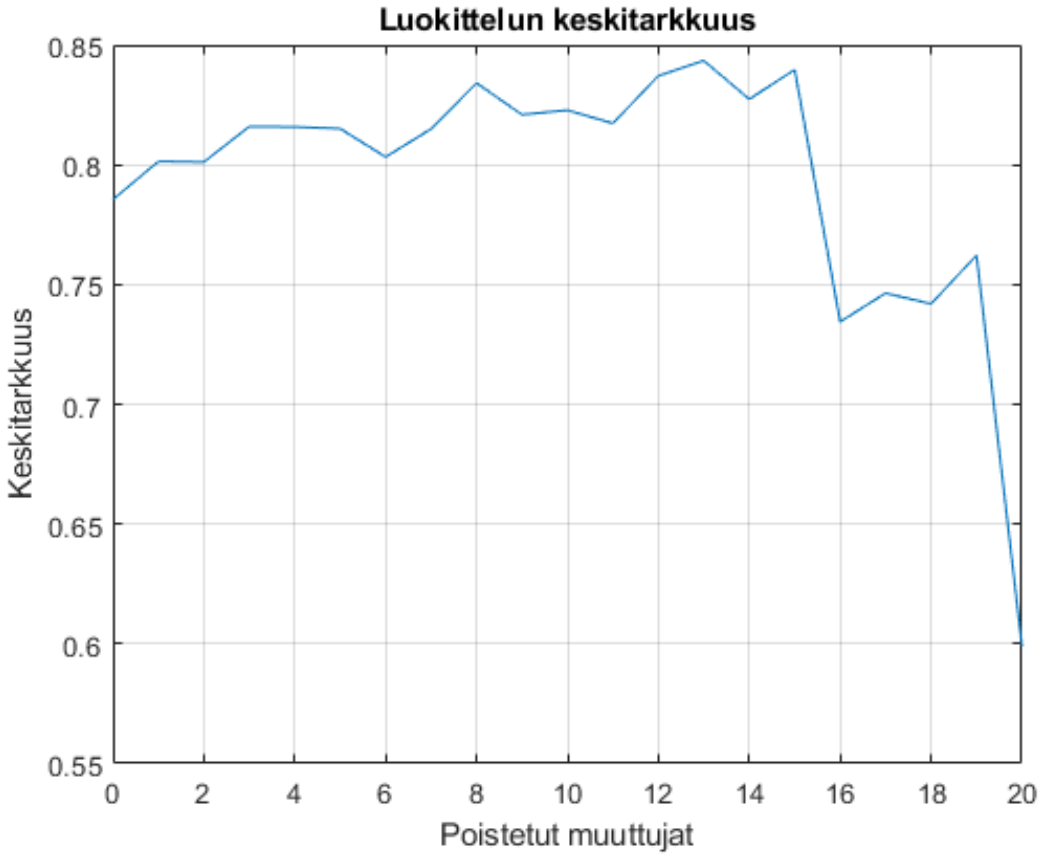
Kuvassa 2 on kuvattu luokittelun keskispesifisyydet eri similaarisuusparametreilla. Kuvaajasta voidaan huomata, että arvot vaihtelevat suuremmin kuvaajan loppupäässä, jolloin saavutetaan myös korkeimmat keskispesifisyyden arvot. Korkein keskispesifisyys 87,57% saadaan similaarisuusarvolla p=2 14:n muuttujan poistamisen jälkeen. Koska similaarisuusarvolla p=0,25 saavutettiin parhaat keskitarkkuuden arvot, tarkastellaan myös sen saamia keskispesifisyyden arvoja. Arvolla p=0,25 saadaan 15:n muuttujan poistamisen jälkeen keskispesifisyydeksi 86,83%, joka ei eroa tilastollisesti merkitsevästi korkeimmasta saadusta arvosta. Koska luokittelun spesifisyys kuvaa luokittimen kykyä tunnistaa negatiiviset tapaukset, voidaan kuvaajasta päätellä, että tietyillä similaarisuusparametrin arvoilla ja muuttujien määrillä negatiiviset tapaukset on luokiteltu hyvin.



Kuva 3. Luokittelun keskiherkkyys muuttujia poistettaessa eri similaarisuusparametreilla

Kuvassa 3 on esitetty luokittelun keskiherkkyydet eri similaarisuusparametreilla. Voidaan havaita, että kuvaaja muistuttaa muodoltaan kuvan 1 kuvaajaa, eli luokittelun keskitarkkuuksia. Keskiherkkyydet ovat arvoiltaan matalampia kuin keskitarkkuudet. Kuvaajien samankaltaisesta muodosta voidaan päätellä, että positiivisia tapauksia luokitellaan negatiivisiksi melko paljon, korkeimman herkkyysarvon ollessa 74,12%. Similaarisuusparametrien p=0,1 ja p=0,25 arvoilla keskiherkkyydet säilyvät selvästi parempina, kun muuttujia poistetaan edelleen.

Eri similaarisuusparametreilla tuotettujen kuvaajien perusteella arvolla p=0,25 tehdyt tulokset ovat mielenkiintoisia ennustemallin kannalta. Tällä parametrilla luokittelun keskitarkkuutta, -spesifisyyttä ja -herkkyyttä saatiin parannettua muuttujanvalinnan avulla 15:n muuttujan poistamiseen saakka.



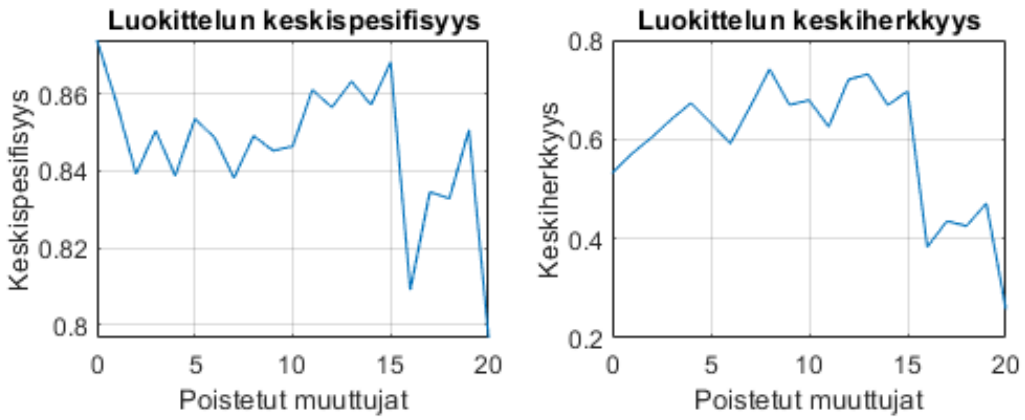
Kuva 4. Luokittelun keskitarkkuus similaarisuusparametrilla 0,25

Kuvassa 4 on esitetty similaarisuusparametrilla p=0,25 saadut luokittelun keskitarkkuudet. Keskitarkkuus nousee 15:n muuttujan poistamiseen saakka. Korkein keskitarkkuuden arvo 84,34% saadaan 13:n muuttujan poistamisen jälkeen. Parhaana arvona voidaan kuitenkin pitää 15. muuttujan poistamisen jälkeen saatavaa 83,97%, sillä arvot eivät eroa toisistaan tilastollisesti merkitsevästi. 16:ntena ja 20:ntena poistetut muuttujat huonontavat luokittelun keskitarkkuutta huomattavasti.

Taulukko 8. Poistetut muuttujat

1	job.admin
2	marital.divorced
3	job.blue-collar/student
4	contact.telephone/cellular/unknown
5	job.technician/services
6	job.self-employed/retired
7	loan.yes/unknown
8	job.management
9	marital.single
10	job.unemployed
11	job.housemaid
12	job.entrepreneur
13	housing.yes/no
14	default.yes/unknown
15	marital.married
16	poutcome.success/other/unknown/failure
17	day
18	pdays
19	month.jan/feb/mar...
20	duration

Taulukossa 8 on esitetty similaarisuusarvoa p=0,25 käyttäen poistettujen muuttujien järjestys. Voidaan huomata, että kaikki asiakkaan työhön ja siviilisäätyyn liittyvät muuttujat poistettiin muuttujanvalinnan alkuvaiheessa, mistä voidaan päätellä, että ne eivät ole ongelman kannalta merkittäviä. Kuvasta 4 nähtiin, että 16:ntena ja 20:ntena poistetut muuttujat vaikuttivat luokittelutulokseen merkittävästi, eli ne ovat ongelman kannalta tärkeitä. 16:s poistettu muuttuja on edellisen kampanjan lopputulos ja 20:s poistettu muuttuja on viime kontaktin kesto.



Kuva 5. Luokittelun keskiherkkyys ja keskispesifisyys similaarisuusparametrilla 0,25

Kuvassa 5 on similaarisuusparametrilla p=0,25 saadut luokittelun keskiherkkyydet ja -spesifisyydet. Nähdään, että keskispesifisyyden arvot ovat korkeampia, eli negatiivisten tapausten luokittelu on onnistunut paremmin. Parhaana luokittelun keskispesifisyyden arvona voidaan pitää 15:n muuttujan poistamisen jälkeen saatavaa 86,83%, sillä se ei eroa tilastollisesti merkitsevästi ennen muuttujanvalintaa saatavasta korkeimmasta arvosta 87,40%. Myös keskiherkkyyden kuvaajasta 15:n muuttujan poistamisen jälkeen saatu arvo 69,69% ei eroa kuvaajan korkeimmasta arvosta 74,12%, joten sitä voidaan pitää parhaana tuloksena. Molemmista kuvaajista on nähtävissä myös 16:nnen ja 20:nnen muuttujien poistamisen vaikutus arvoihin.

Similaarisuusarvolla p=0,25 15:n muuttujan poistamisen jälkeen saadaan ennustemalli, jota voidaan pitää parhaana, koska se on suoriutunut hyvin sekä tarkkuudeltaan, herkkyydeltään että spesifisyydeltään. Lisäksi siinä on jäljellä vain kymmenen muuttujaa, eli laskennan määrä on vähentynyt huomattavasti. Jäljellä olevat muuttujat ovat viime kampanjan lopputulos, kampanjan aikaisten yhteydenottojen määrä, keskimääräinen vuotuinen saldo, koulutuksen taso, ikä, viime kampanjan lopputulos, viime kontaktista kuluneet päivät, viime kontaktin päivä, viime kontaktin kuukausi ja viime kontaktin kesto. Ennustemallin saavuttama keskitarkkuus on 83,97%, keskispesifisyys 86,83% ja keskiherkkyys 69,69%. Taulukossa 9 on vertailtu ennustemallilla saatuja tuloksia alkutilanteeseen.

Taulukko 9. Tulokset ennustemallilla ja alkutilanteessa

	Ennustemalli	Alkutilanne
Varianssi	0,0001109	0,001130
Keskihajonta	0,0105	0,0336
Keskiarvotarkkuus 99% luottamusvälillä	$0,8397 \pm 0,0045$	$0,7854 \pm 0,0145$
Keskiarvospesifisyys 99% luottamusvälillä	$0,8683 \pm 0,0045$	$0,8740 \pm 0,0145$
Keskiarvoherkkyys 99% luottamusvälillä	$0,6969 \pm 0,0045$	$0,5321 \pm 0,0145$
Optimaalinen p	3,75	2,25
Optimaalinen m	2	3,75

Tutkitaan seuraavaksi luokittelun hyödyllisyyttä yksinkertaisen simulaation avulla. Tarkastellaan massa- ja suoramarkkinointikampanjoiden tulosten eroa. Simuloidaan tilannetta, jossa keskimääräinen asiakaskontaktin kesto on 12 minuuttia ja markkinoiminen maksaa 18 euroa tunnissa. Keskimääräinen asiakkaan tekemä talletus on 800 euroa, josta pankki tekee voittoa 7% sijoitettuaan talletuksen. Taulukossa 10 on esitetty simuloitu tilanne, jossa eri virheluokkille on laskettu erikseen kulut ja mahdollinen voitto.

Taulukko 10. Simulaatio luokittelun mukaisesta markkinoinnista

Virheluokka	Puheluiden määrä	Käytetty aika (h)	Kulut (e)	Tehdyt talletukset (e)	Talletuksilla saatu voitto (e)
FN	9	1,8	32,4	7200	504
FP	30	6	108	0	0
TN	169	33,8	608,4	0	0
TP	21	4,2	75,6	16 800	1176

Taulukon 10 mukaan laskettuna massakampanjan tulos on 855,60 euroa ja kohdennetun kampanjan tulos 992,40 euroa. Luokittimen mukaan tehtyä kohdennettua kampanjaa käyttäen tulos on 16,0% korkeampi. Luotua ennustemallia hyödyntäen toteutettu markkinointikampanja on siis merkittävästi tuottavampi kuin massakampanja.

5 KESKUSTELU

Erilaisia muuttujanvalintamenetelmiä vertailtaessa yksi tärkeä tekijä on menetelmän laskennallinen vaativuus, jonka vuoksi suodatinmenetelmä soveltui ratkaisun luomiseen hyvin. Ennustemallia luotaessa huomattiin, että laskenta-aika kasvoi huomattavasti, kun dataa jaettiin useita kertoja testaus- ja opetusdataan. Tämän vuoksi on tärkeää, että luotava ennustemalli ei ole liian monimutkainen ja laskennallisesti aikaavievä. Eri menetelmällä oltaisiin voitu päästä tarkempiin tuloksiin, mutta mallista olisi tullut hitaampi.

Parhaaksi nähty ennustemalli hyödynsi muuttujina viime kampanjan lopputulosta, kampanjan aikaisten yhteydenottojen määrää, keskimääräistä vuotuista saldoa, koulutuksen tasoa, ikää, viime kampanjan lopputulosta, viime kontaktista kuluneet päivät, viime kontaktin päivää, viime kontaktin kuukautta sekä viime kontaktin kestoa. Tästä voidaan päätellä, että asiakkaan työhön tai siviilisäätyyn liittyvät muuttujat olivat ongelman kannalta merkityksettömiä ja kampanjaan liittyvät muuttujat olivat selkeästi tärkeämpiä. Viime kampanjan lopputulos ja viime kontaktin kesto nousivat esille tuloksia tarkasteltaessa erityisen tärkeinä muuttujina. Markkinointidataa kerätessä kampanjaan ja markkinointitapaan liittyvän tiedon kerääminen voikin olla oleellisempaa kuin asiakkaaseen liittyvän tiedon saaminen.

6 JOHTOPÄÄTÖKSET

Työn tarkoituksena oli luoda ennustemalli, jonka avulla voidaan ennustaa, kannattaako tietylle kohderyhmälle lähteä markkinoimaan. Aihetta lähestyttiin epämääräisyyteen perustuvan muuttujanvalintamenetelmän ja similaarisuuteen pohjautuvan luokittelualgoritmin kautta. Eri similaarisuusarvoilla luokiteltuja tuloksia tarkastellessa löydettiin ongelmaan sopiva ennustemalli.

Huomattiin, että datan esikäsittely on tärkeässä osassa ennustemallin luomisessa. Sopivalla similaarisuusparametrilla löydetty malli onnistui parantamaan luokittelutulosta viidentoista muuttujan poistamiseen saakka. Päästiin siis parempiin luokittelutuloksiin sekä onnistuttiin yksinkertaistamaan ennustemalli käyttämään kymmentä muuttujaa. Luokittelun keskitarkkuudeksi saatiin 83,97%.

Yksi ongelman kannalta tärkeä näkökulma on eri virheluokat, joista false negative nähtiin vakavimpana. Luokittelun herkkyyden avulla pystyttiin tutkimaan, kuinka hyvin positiiviset tapaukset on tunnistettu. Saavutettiin keskiherkkyys 69,69%, eli positiiviset tapaukset tunnistettiin melko hyvin. Sen sijaan luokittimen keskispesifisyys oli 86,83%, eli negatiiviset tapaukset tunnistettiin paremmin kuin positiiviset. Tulevaisuudessa mallia voitaisiin kehittää niin, että se tunnistaisi positiiviset tapaukset paremmin tai käyttää siihen paremmin soveltuvaa menetelmää.

Lähteet

[1] Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2009.

[2] Blum, Avrim L ja Langley, Pat. "Selection of relevant features and examples in machine learning". *Artificial intelligence* 97.1-2 (1997), s. 245–271.

[3] Cardie, Claire. "Using decision trees to improve case-based learning". Teoksessa: *Proceedings of the tenth international conference on machine learning*. 1993, s. 25–32.

[4] Dash, Manoranjan ja Liu, Huan. "Feature selection for classification". *Intelligent data analysis* 1.1-4 (1997), s. 131–156.

[5] Duda, Richard O, Hart, Peter E ja Stork, David G. *Pattern classification*. John Wiley & Sons, 2012.

[6] Goldberg, David E. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN: 0201157675.

[7] Guyon, Isabelle ja Elisseeff, André. "An introduction to variable and feature selection". *Journal of machine learning research* 3.Mar (2003), s. 1157–1182.

[8] John, George H, Kohavi, Ron ja Pfleger, Karl. "Irrelevant features and the subset selection problem". Teoksessa: *Machine Learning Proceedings 1994*. Elsevier, 1994, s. 121–129.

[9] Kira, Kenji ja Rendell, Larry A. "The feature selection problem: Traditional methods and a new algorithm". Teoksessa: *Aaai*. Vol. 2. 1992, s. 129–134.

[10] Klir, George J ja Yuan, Baozung. *Fuzzy sets and fuzzy logic: theory and applications*. Vol. 574. Prentice Hall PTR New Jersey, 1995.

[11] Kohavi, Ron ja John, George H. "Wrappers for feature subset selection". *Artificial intelligence* 97.1-2 (1997), s. 273–324.

[12] Koller, Daphne ja Sahami, Mehran. *Toward optimal feature selection*. Tekninen raportti. Stanford InfoLab, 1996.

[13] Luukka, Pasi ja Lohrmann, Christoph. “Information transmission and nonspecificity in feature selection”. *Proceedings of 2019 IFSA World Congress and NAFIPS Annual Conference* (2019).

[14] Luukka, Pasi, Saastamoinen, Kalle ja Kononen, Ville. “A classifier based on the maximal fuzzy similarity in the generalized Lukasiewicz-structure”. Teoksessa: *10th IEEE International Conference on Fuzzy Systems.(Cat. No. 01CH37297)*. Vol. 1. IEEE. 2001, s. 195–198.

[15] Moro, S., Laureano, R. ja Cortez, P. “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology”. Teoksessa: *Proceedings of the European Simulation and Modelling Conference - ESM’2011*. Toim. P. Novais et al. Guimaraes, Portugal: EUROSIS, lokakuu 2011, s. 117–121.

[16] Narendra, Patrenahalli M. ja Fukunaga, Keinosuke. “A branch and bound algorithm for feature subset selection”. *IEEE Transactions on computers* 9 (1977), s. 917–922.

[17] Tan, Ding-Wen, Sim, Yee-Wai ja Yeoh, William. “Applying feature selection methods to improve the predictive model of a direct marketing problem”. Teoksessa: *International Conference on Software Engineering and Computer Systems*. Springer. 2011, s. 155–167.

[18] Yu, Lei ja Liu, Huan. “Feature selection for high-dimensional data: A fast correlation-based filter solution”. Teoksessa: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, s. 856–863.

[19] Zadeh, Lotfi A. “Similarity relations and fuzzy orderings”. *Information sciences* 3.2 (1971), s. 177–200.

Taulukot

Kuvat